7.1 A 4b/cell NROM 1Gb Data-Storage Memory

Yan Polansky[1], Avi Lavan[1], Ran Sahar[1], Oleg Dadashev[1], Yoram Betser[1], Guy Cohen[1], Eduardo Maayan[1], Boaz Eitan[1], Ful-Long Ni[2], Yen-Hui Joseph Ku[2], Chih-Yuan Lu[2], Tim Chang-Ting Chen[2], Chun-Yu Liao[2], Chin-Hung Chang[2], Chung-Kuang Chen[2], Wen-Chiao Ho[2], Yite Shih[2], Wenchi Ting[2], Wenpin Lu[2]

[1]Saifun Semiconductors, Netanya, Israel
[2]Macronix, Hsinchu, Taiwan

The use of non-volatile memories for mass storage applications is mainly cost driven. One cost-reduction technique is having 2b or even better 4b per cell manufactured in a low-cost process. In this paper, a 1Gb data flash device based on 4b/cell NROM technology is presented.

The NROM cell is based on localized charge trapping in an ONO gate dielectric [1]. When a standard NROM cell is operated in a 4b mode, two bits of data are stored in each of its two separated storage areas (Fig. 7.1.1). The inherent advantage of the 4b NROM is that only 4 V_t levels are required in each storage area.

The key challenges in developing a 4b NROM product are: 1) accurate and fast programming algorithm, and 2) reliable and fast sensing scheme.

The V_t distributions of the programmed states in a 4b NROM product must be narrow and well controlled. A two-phase programming algorithm is implemented to provide a fast first phase, and a slower, but more accurate second phase. The first phase is based on a drain-stepping algorithm for reaching a V_t that is lower than the final target while the second phase is based on a gate-stepping algorithm. This two-phase algorithm fullfils the accuracy requirements, but requires detailed optimization to meet data flash programming rate and reliability requirements. In the 4b NROM 1Gb device this optimization is possible and simple due to the flexibility of software-based algorithms being used. The algorithms are operated by a simple embedded 8b microcontroller, where its microcode is stored in an NROM-based 4k-word programmable ROM (PFROM). The programming algorithm is executed on multiple cells in parallel taking into account the performance requirements, the programming level, the charge pump capacity, and more. The typical programming speed in a 4b NROM device using the two-phase algorithm is 3.5MB/s. A typical V_t distribution plot measured on the 1Gb device is shown in Fig. 7.1.2.

The 4b NROM 1Gb device has a segmented virtual-ground (VG) array architecture (Fig. 7.1.3) similar to the one used in 2b NROM products [2, 3]. The main difference between the 2b and 4b arrays is the select transistors (SEL) scheme. In the 4b NROM array, the SEL transistors are configured to enable simultaneous access to three local BLs within a slice through three different global BL (GBL). Array access through 3 GBLs is required in order to implement a "neighbor effect" cancellation scheme. Other differences, which are dictated by performance and specification parameters are the number of WLs (512) and the number of NROM cells (35) within the basic VG slice.

The 4b NROM 1Gb device uses a drain-side ac-sensing scheme. Drain-side sensing is preferred over source-side sensing [2, 3] in order to better suppress the "neighbor effect", i.e., the leakage current flowing through the VG pipe during read. When accessing a cell at read, an integrating capacitor (coupled to the drain side of the cell) is used to generate the read signal of the cell. A large read signal is obtained by decoupling the integrating capacitor from the GBL capacitance. The cell signal is then compared to a global dc level (Vcomp) by a sense amplifier (SA) having an offset-cancellation scheme. A set of three reference cells is used to readout in parallel the 2b stored in each accessed storage area. In order to match the operating window of all the cells, a single set of references is used per half array. The operation of the reference cells is matched to the array cells operation, but the outputs of the SAs of the reference cells are used as latching signals to capture the array cells data, creating a time-domain sensing scheme. A block diagram of the sensing scheme of the half array (32 SAs) is shown in Fig. 7.1.4, and the main signals diagram is shown in Fig. 7.1.5. In order to meet the required read performance, 64 SAs are operated in parallel to readout 128b of data at once, leading to a >20MB/s read rate.

The sensing scheme incorporates a simple error-detection (ED) scheme used to detect if the data read out from the array is identical to the data initially inserted by the user. The ED is accomplished by automatic attachement of additional bits to each chunk of data being written to the array. These additional ED bits can provide the number of bits per level in each chunk of data. Then, having this ED data available, a read error can be detected if while reading out the array data, it is counted and compared to the stored ED bits data. In order to improve the ED bits reliability, they are written to the array in a 2b/cell mode. Therefore, each page contains 4b NROM cells storing the user data, and 2b NROM cells storing the ED data.

In order to reduce the effects of the statistical variations of cells parameters and characteristics within a full page, the page is segmented into chunks of data and ED bits are attached to each of these chunks.

In NROM technology, after cycling and retention, the entire distributions move rigidly with no single bit failures. This is a fundamental feature of the NROM technology and as long as a window remains between states, all the cells can be reliably read and all levels can be accurately resolved. The only requirement is the ability to correctly locate the read reference level within the remaining window between the V_t states. This is illustrated in Fig. 7.1.6, where V_t distributions before and after cycling/retention can be correctly read if the read reference is located for example at RD1 and RD2 levels, respectively. The choice of the most appropriate read reference level is based on the ED results. When a page is accessed for read, the first chunk is read out and all the levels are evaluated with reference to the ED bits of that chunk. If an error is detected, then the read reference levels are moved (by adjusting the reference cells WL level), and the same chunk of data is re-read. This continues until the chunk of data is correctly read. The next chunks of the page are then read out following a similar procedure. The number of repeated reads allowed and the data chunk size determine the first-byte-latency performance of the device. In the 1Gb 4b NROM device the 2kB page data are segmented to chunks of 128B, each with its own ED bits, reaching a first-byte read-access latency of 15µs.

The 4b NROM based 1Gb device is manufactured in a standard low-cost 0.13µm NROM process, having a single gate oxide (i.e., only high-voltage CMOS), 2 metals, and achieving an 85mm^2 die area. The device and process specifications are listed in Figure 7.1.7. Figure 7.1.8 shows a micrograph of the 1Gb 4b/cell NROM data storage device die with its major blocks marked.

The integration of a fast and accurate sensing scheme, a simple ED concept, and a moving-reference scheme, all operated by an embedded microcontroller, makes 4b NROM products a reality.

References:

[1] B. Eitan, et al., EDL-21 (11), pp. 543-545, 2000.
[2] E. Maayan, et al., "A 512Mb NROM Data Storage Memory with 8MB/s Data Rate," *ISSCC Dig. Tech. Papers*, pp. 100-101, Feb., 2002.
[3] Y. Sofer, et al., "A 55mm^2 256Mb NROM Flash Memory with Embedded Microcontroller Using an NROM-Based Program File ROM," *ISSCC Dig. Tech. Papers*, pp. 48-49, Feb., 2004.

Figure 7.1.1: The 4b NROM.

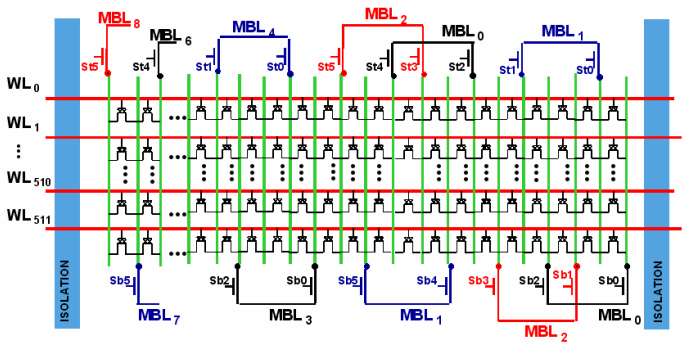

Figure 7.1.2: The virtual-ground array slice.

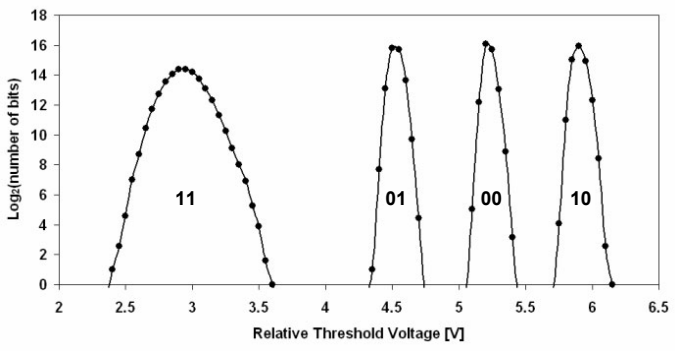

Figure 7.1.3: Vt distributions measured on the 4b NROM 1Gb device.

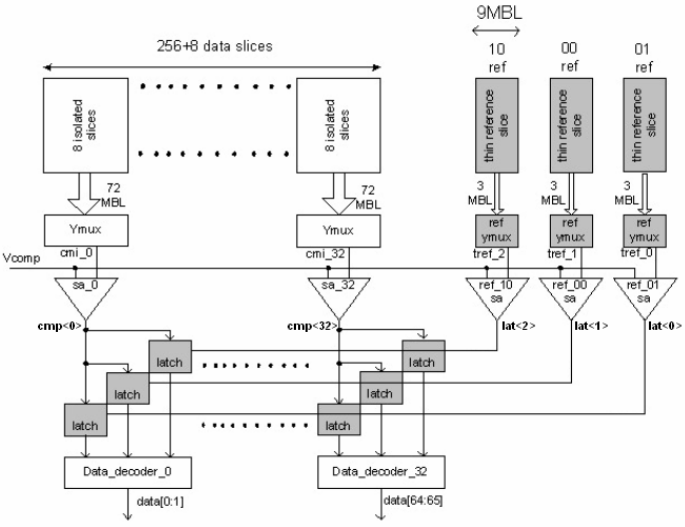

Figure 7.1.4: Block diagram of the time-domain multilevel sensing scheme.

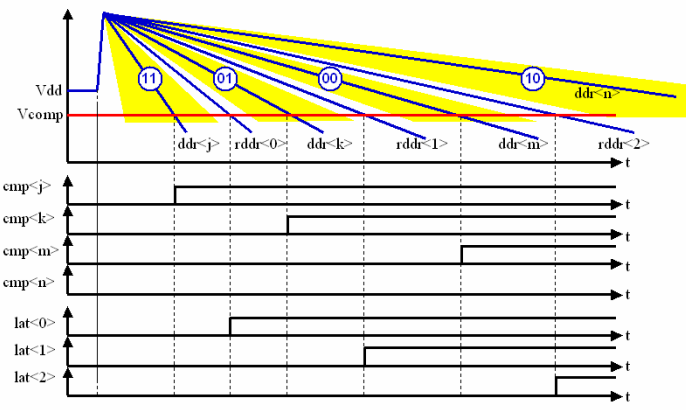

Figure 7.1.5: Sense amplifiers input and output signals.

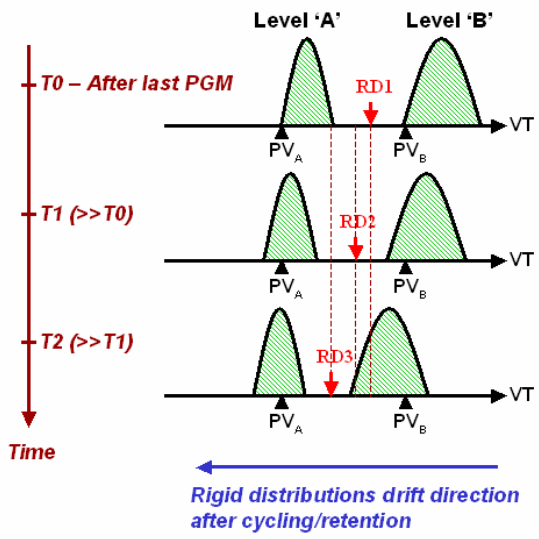

Figure 7.1.6: The moving read-reference concept.

Continued on Page 644

Parameter	Unit	Value
Effective WL Width	[μm]	0.13
Effective BL Width	[μm]	0.2
Cell Size	[F²/Cell]	5
4b MLC Array Cell Density	[F²/b]	1.25
4b MLC Bit Size	[μm²]	0.021

Figure 7.1.7: Device and process specification table.

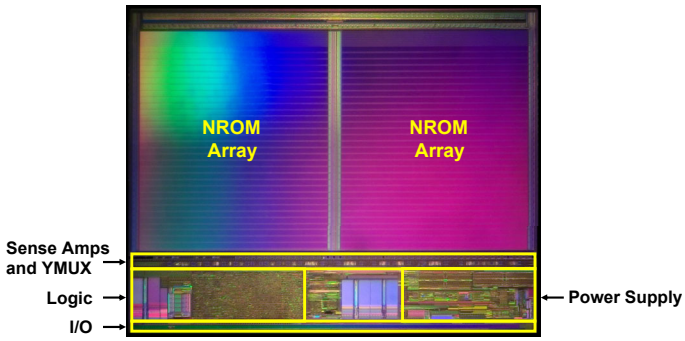

Figure 7.1.8: Die micrograph.

